

# Relating Diversity and Human Appropriation from Land Cover Data <sup>\*</sup>

Carme Font  
Department of Mathematics  
Universitat Autònoma de Barcelona  
08193 Bellaterra, Catalonia  
carmefont@mat.uab.cat

Mercè Farré  
Department of Mathematics  
Universitat Autònoma de Barcelona  
08193 Bellaterra, Catalonia  
farre@mat.uab.cat

Aureli Alabert  
Department of Mathematics  
Universitat Autònoma de Barcelona  
08193 Bellaterra, Catalonia  
Aureli.Alabert@uab.cat

October 25, 2016

## Abstract

We present a method to describe the relation between indicators of landscape diversity and the human appropriation of the net primary production in a given region. These quantities are viewed as functions of the vector of proportions of the different land covers, which is in turn treated as a random vector whose values depend on the particular small terrain cell that is observed.

We illustrate the method assuming first that the vector of proportions follows a uniform distribution on the simplex. We then consider as starting point a raw dataset of observed proportions for each cell, for which we must first obtain an estimate of its theoretical probability distribution, and secondly generate a sample of large size from it. We apply this procedure to real historical data of the Mallorca Island in three different moments of time.

Our main goal is to compute the mean value of the landscape diversity as a function of the level of human appropriation. This function is related to the so-called Energy-Species hypothesis and to the Intermediate Disturbance Hypothesis.

**Keywords:** Diversity, Net Primary Production, Human Appropriation, Mallorca Island, Compositional Data, Dirichlet Distribution, Estimation of Densities, Simulation.

**Mathematics Subject Classification (2010):** 62P12, 62G07, 65C10

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Uniform distribution of land covers</b>	<b>5</b>
2.1	The distribution of $H$	6
2.2	The distribution of $A$	6
2.3	Expected value of $H$ for a given appropriation	8

---

<sup>\*</sup>This work has been partially supported by grant number UNAB10-4E-378, co-funded by the European Regional Development Fund (ERDF); grant number HAR2015-69620-C2-1-P funded by MINECO, and the International Partnership Grant SSHRC- 895-2011-1020, funded by the Social Sciences and Humanities Research Council of Canada.

<b>3</b>	<b>Shannon index and appropriation with real data</b>	<b>9</b>
3.1	Kernel density estimation on the simplex . . . . .	11
3.2	Numerical approximation of the estimated density . . . . .	12
3.3	Sampling strategy . . . . .	13
3.4	Choosing the bandwidth parameter . . . . .	14
<b>4</b>	<b>Results</b>	<b>15</b>
4.1	Shannon index conditioned to the appropriation . . . . .	16
4.2	Diversity and urban land cover . . . . .	17
<b>5</b>	<b>Computational notes</b>	<b>18</b>

## 1 Introduction

The *Net Primary Production* (NPP) is the net amount of solar energy converted to plant organic matter through photosynthesis. The *Human Appropriation of Net Primary Production* (HANPP) is an indicator of the alterations produced by human activity on the NPP (see, for instance, [19], [11], [10]). These alterations include the degradation of the environment (which leads to differentiate between the potential NPP and the actual NPP) and the harvesting of photosynthetic products, which further reduces the actual NPP to a quantity sometimes denoted  $NPP_t$ . Thus,  $HANPP = NPP_{pot} - NPP_t$ .

It is customary to measure the human appropriation as a percentage of the potential primary production:  $HANPP\% = 100 \times HANPP/NPP_{pot}$ . One way to approximate the HANPP% of a given area is to assign a coefficient  $w_i$  to each of the  $n$  different land uses present in the area and compute the weighted average

$$HANPP\% = \sum_{i=1}^n w_i p_i, \quad (1)$$

where  $p_i$  are the proportions of land devoted to each use. The weights  $w_i$  indicate the percentage of human appropriation for each specific land use. We will speak more generally of *land covers* (forest, wetlands, crop, etc.).

Ideally, we would like to relate human appropriation with some measure of the biodiversity in a given agro-ecosystem, in order to assess how human activity affects other species.

There are several indices aimed at measuring biodiversity. The most popular one is the *Shannon index*, defined by the entropy formula

$$H = - \sum_{k=1}^s q_k \log q_k,$$

where  $q_i$ ,  $i = 1, \dots, s$  is the proportion of each of the  $s$  species of a certain group which are present in a certain ecosystem.

The Shannon index is sensitive both to the species richness and to its evenness in the following precise sense: If, for some  $j$ ,  $0 \leq q_j < q_k$  holds for all  $k \neq j$ , then a small increase in  $q_j$  without increasing any of the other proportions results in an increase of  $H$ . The base of the logarithm is arbitrary; if we take base  $s$ , then  $0 \leq H \leq 1$ .

The *Simpson diversity index*

$1 - \sum_{k=1}^s q_k^2$  also increases with species richness and evenness, in the same sense above, whereas the less used *Berger–Parker index*  $(\max q_i)^{-1}$  is only sensitive to the proportion of the most populated species. These indices, and some others, can be seen as particular cases of a family of measures (see [13]).

When we say “number of species” we are of course talking of a given taxonomic group of living organisms (such as birds, butterflies, trees, insects, mammals, herbivores, carnivores, primary producers,

...), possibly grouping together similar species. Obtaining an actual biodiversity index in a given region by direct observation and sampling is very difficult [4].

Suppose anyway that we have a good estimate of the proportion of species in each particular land cover. Assume that we have  $n$  different land covers coexisting in a given area in proportions  $p_i$ ,  $i = 1, \dots, n$ . Let  $s$  be the total number of species in the area, and  $q_{ik}$  the proportion of species  $k$  in cover  $i$ . Then the Shannon index of the area is

$$-\sum_{i=1}^n \sum_{k=1}^s q_{ik} p_i \log(q_{ik} p_i) . \quad (2)$$

In this formula we are assuming that species living in different covers are considered different, thus in fact it combines bio- and land-cover-diversity. Eventually, the proximity of some covers may produce the appearance of new species that are not present when the covers are not close (see again [4]).

According to the so-called *species-energy hypothesis* (see e.g. the survey by [5] on this and other hypotheses, and the references therein), the richness of species is monotonically increasing as a function of the available energy in the system. This would explain, for example, the richness gradient from the poles to the tropics, as the energy provided by the sun is greater at lower latitudes.

At geographical (large) scales, it has been suggested that this is true through all energy levels, although there is still little empirical evidence in this generality. At local scales this is not at all clear, and [7], among others, writes that “there is a marked tendency for a general hump-shaped relationship between species richness and available energy”. In other words, that whereas when the available energy goes from low to moderate levels, richness indeed increase, from moderate to high levels the relation is reversed.

In terms of HANPP, which represents energy that humans take out of the natural system, Gaston’s remark amounts to say that biodiversity, as a function of HANPP, increases at the beginning, peaks at a certain point, and then decrease again when HANPP is high. The empirical work of [11], who measured the number of species of 9 groups (plant and animal) on 38 small Austrian regions of similar characteristics, confirms that above 40-50% of total possible HANPP, species richness indeed decreases, but there are no data below these percentages. The authors adjust a linear decreasing relationship, although graphically the decrease seems to be more “concave” than linear in most cases.

The possibility that low values of HANPP lead to diversity values below the maximum seems to be related to the so-called *Intermediate Disturbance Hypotheses* (IDH), which states that moderate disturbances or fluctuations of any kind in an environment lead to more diversity than strong or weak disruptions. It should be remarked that IDH is controversial, as it is the species-energy hypothesis. For instance, the recent review article by [6] is clearly against. In any case, the intermediate disturbance in natural systems should be understood as punctual interventions or catastrophes, whereas in an agro-cultural system it is the result of the continuous human intervention.

Numerous studies haven been published relating landscape heterogeneity with biodiversity. [18] contains a large review of articles on this subject; in most of them it is concluded that landscape diversity is positively correlated with species diversity. With this fact in mind, and taking into account the difficulty to evaluate the biodiversity of a given area, we will use the Shannon index relative to land covers

$$H = -\sum_{i=1}^n p_i \log p_i \quad (3)$$

as our measure of ‘diversity’. The Simpson and Berger-Parker indices could be redefined in the same way, replacing species proportions  $q_k$  by land cover proportions  $p_i$ . Strictly speaking, however, (3) is only an indicator of the degree of mosaic structure of a piece of land; [16] use a combination of  $H$  and a so-called Ecological Connectivity Index to model biodiversity.

In this paper we try to relate human appropriation as defined by (1) with the Shannon entropy index given by (3). For the sake of brevity, we will denote the HANPP% measure of (1) simply by  $A$  in the formulae throughout the paper, and we will speak of (human) *appropriation*.

Both  $H$  and  $A$  are functions of the land proportions  $p_i$  in a terrain cell, but we would like somehow to obtain a “function” yielding  $H$  from the appropriation alone. Actually, this is not possible in a strict sense, since the same value of appropriation may correspond to many values of entropy, and vice versa. We propose the following setup:

On a given terrain cell  $\omega$ , the different land covers may appear in certain proportions  $0 \leq p_i(\omega) \leq 1$ . Suppose we observe a big number of such cells, and we apply a fixed set of coefficients  $w_1, \dots, w_n$  to all of them.

Then, cell  $\omega$  has a certain appropriation  $A(\omega)$  and a certain entropy  $H(\omega)$ . We may think that  $\omega$  is a random parameter, so that  $p_i(\omega)$  are random proportions and also  $A(\omega)$  and  $H(\omega)$  are random. We aim at describing the probability distribution of  $H$  given a certain value  $A(\omega) = a$  of the appropriation, for every possible  $a$ . It is therefore the probability distribution of  $H$  which will be a function of  $A$ .

Our case study is *Mallorca*, a Mediterranean island with a total area of 3,603 km<sup>2</sup> of calcareous origin. The mountain range of *Serra de Tramuntana* runs parallel to the North coast and reaches 1,445 metres in the highest peak. Between this range and the eastern mountains of *Serres de Llevant*, a plain occupies most of the island. Annual precipitation varies from 300 mm (in the South) to 1,800 mm (in the North) with an average temperature of 16 °C. We work with land cover data based on land cover maps of Mallorca (Figure 1) obtained from [8] for three time periods (1956, 1973, 2000). These data comprises a total of 3360 cells of size  $1 \times 1$  km<sup>2</sup>, once disregarded those with some part into the sea.

We have grouped land covers into four categories, namely ‘semi-natural’, ‘croplands’, ‘groves’ and ‘urban’. Semi-natural land covers include forest, scrub, prairie and bedrock, and wetlands. Croplands include both dry and irrigated croplands. Groves are composed of rain-fed arboricultural groves, irrigated groves and olive groves. Urban land covers are both urban and industrial areas.

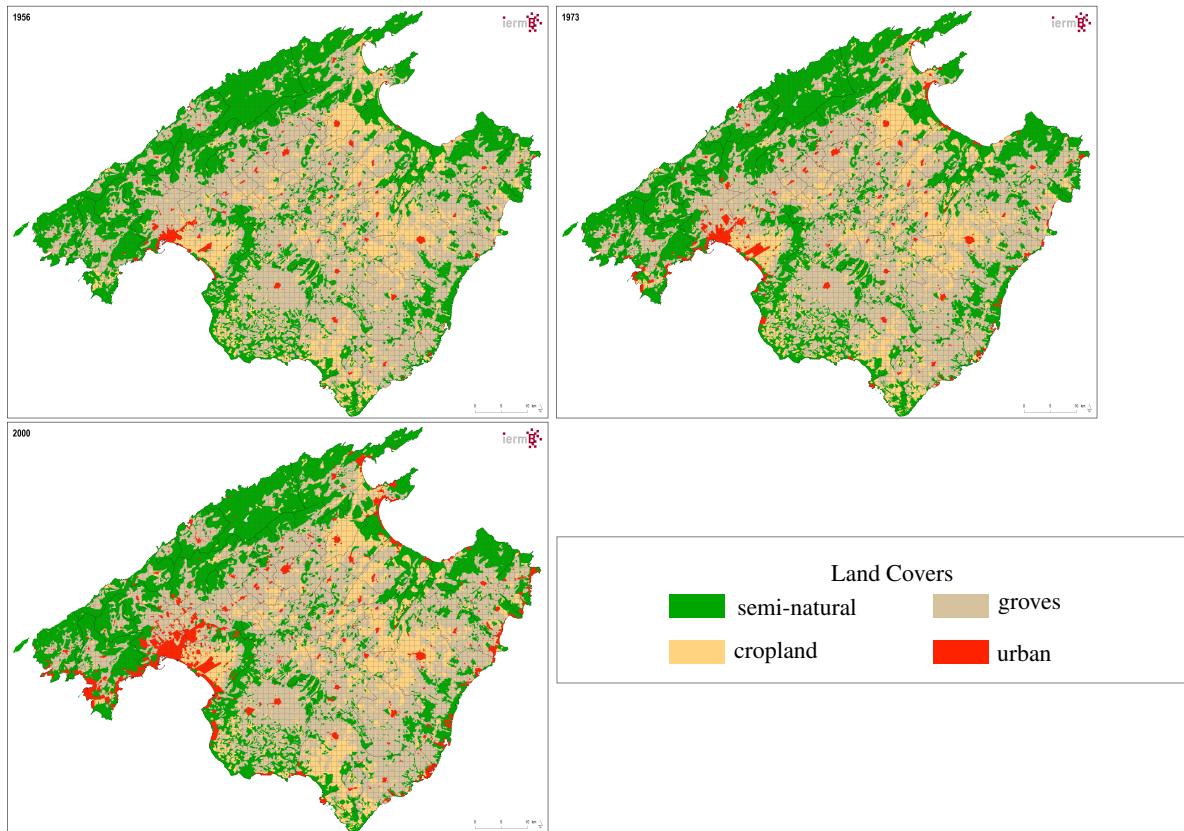


Figure 1: Mallorca land cover maps at regional scale (1:50,000) for 1956, 1973 and 2000. Source: from [8] in collaboration with the Barcelona Institute of Regional and Metropolitan Studies.

The methodology proposed here to relate the Shannon entropy  $H$  with the appropriation  $A$  can also

be used with the other indices of diversity cited in this introduction, or with other functions of land cover proportions. The human appropriation can be either a random variable whose distribution is determined by a given theoretical distribution of land-covers (case treated in Section 2), or a function of empirically obtained data (developed in Section 3, with our case study in mind).

In Section 2, we assume a simple uniform probability distribution of proportions of land covers and

- a) we show how to obtain (by simulation) the distribution of the entropy  $H$ , and we compute (exactly) its expected value;
- b) we compute the distribution of the appropriation  $A$  and its expectation, and
- c) we derive a formula for the conditional expectation of  $H$  given any fixed value of the appropriation.

In Section 3, we estimate the conditional expectation of  $H$  given  $A$  using real sample data. This involves estimating the probability distribution from which the data has been (ideally) originated, and produce a very large sample following the estimated distribution. The process has some difficulties which are explained at the beginning of the section, and developed in several subsections.

Finally, some specific data-related details and the results of the case study are presented in Section 4.

## 2 Uniform distribution of land covers

Given a set of cells  $\Omega$ , and a set of  $n + 1$  possible land covers, we have defined the appropriation and Shannon indices of each cell  $\omega \in \Omega$ , by

$$\begin{aligned} A(\omega) &= \sum_{i=1}^{n+1} w_i p_i(\omega) \\ H(\omega) &= - \sum_{i=1}^{n+1} p_i(\omega) \log_{n+1} p_i(\omega) \end{aligned} \tag{4}$$

where  $p_i(\omega)$  is the proportion of cover  $i$  in cell  $\omega$ , and we arbitrarily take  $n + 1$  as the base of the logarithm, so that the maximal value that  $H$  can achieve is normalised to 1. Working in dimension  $n + 1$  instead of  $n$  simplifies the notation later.

We study the relation between these two quantities by postulating some probability distribution of the random vector  $p(\omega) = (p_1(\omega), \dots, p_{n+1}(\omega))$ . Notice that this vector takes values in the so-called *standard  $n$ -simplex in  $\mathbb{R}^{n+1}$* , i.e. the  $n$ -dimensional surface

$$\Delta = \{(p_1, \dots, p_{n+1}) \mid p_i \geq 0, p_1 + \dots + p_{n+1} = 1\} .$$

We are thus working with *compositional data* (see e.g. [2]).

In this section we will assume that  $p$  follows the uniform distribution on the simplex. This assumption does not aim to represent any realistic situation; for instance, it implies that all covers are actually present in some proportion in all cells. But it is anyway the usual modelling choice when no other information is present.

The volume of the standard  $n$ -simplex is  $\frac{\sqrt{n+1}}{n!}$ , whence the density of the uniform distribution is given by

$$f(p_1, \dots, p_{n+1}) = \begin{cases} \frac{n!}{\sqrt{n+1}} & \text{if } p \in \Delta \\ 0 & \text{otherwise} . \end{cases}$$

The marginal distribution of the first  $n$  coordinates is also uniform, on the projected simplex

$$\Delta' = \{(p_1, \dots, p_n) \mid p_i \geq 0, p_1 + \dots + p_n \leq 1\} ,$$

with the density

$$f(p_1, \dots, p_n) = \begin{cases} n! & \text{if } p \in \Delta' \\ 0 & \text{otherwise} \end{cases}.$$

We can easily obtain the marginal density function of  $p_i$  integrating  $f$  with respect to  $p_j$ ,  $j \neq i$ . For  $p_1$ ,

$$\begin{aligned} f(p_1) &= n! \int_0^{1-p_1} \dots \int_0^{1-\sum_{i=0}^{n-1} p_i} dp_n \dots dp_2 \\ &= n(1-p_1)^{n-1}, \end{aligned} \quad (5)$$

and by symmetry the same formula holds for all  $p_i$ .

It is better to work in the projected simplex, since  $f$  is then a true density with respect to Lebesgue measure in  $\mathbb{R}^n$ , whereas on the standard simplex the support of the probability has zero measure as a subset of  $\mathbb{R}^{n+1}$ .

In the next subsections we study the probability distribution of the random variables  $H$  and  $A$ , and the conditional expectation of  $H$  given  $A$ . We will in general avoid to write explicitly the random parameter  $\omega$  from which the land covers depend.

## 2.1 The distribution of $H$

It is not possible to find analytically the probability distribution of  $H$  from the law of  $p$ . However it is trivial to generate random samples of  $p$  according to the uniform distribution on the simplex and draw a histogram of values of  $H$  using (4). In Figure 2, we show those histograms for 3 and 4 land covers, obtained with a sample size of one million. We have also added to the figure an estimation of the density function of  $H$  and the position of the sample mean.

The density estimation has been carried out using the logsplines method implemented in the R package `logspline` [12]. The usual kernel methods to estimate densities are not suitable here because  $H$  is a bounded random variable. The base uniform sample on the simplex has been generated using the algorithm explained in [17]: If  $Y_1, \dots, Y_{n+1}$  are independent unit-exponential random variables, and

$$E_i = \frac{Y_i}{\sum_{j=1}^{n+1} Y_j}, \quad (6)$$

then the random vector  $(E_1, \dots, E_{n+1})$  is uniformly distributed on  $\Delta$ .

In fact, one does not need to estimate the theoretical mean of the distribution of  $H$  by simulation, since it can be computed exactly. Indeed,

the integral of  $x \log_{n+1} x$  against the density (5), yields

$$\frac{-1}{\ln(n+1)} \left[ \frac{\Psi(n+1) + \gamma - 1}{n+1} + \frac{1}{(n+1)^2} \right],$$

where  $\Psi$  is the digamma function, and  $\gamma$  is the Euler-Mascheroni constant. Therefore, the expectation of the Shannon index  $H$  of (4), under the hypothesis of uniform distribution of the proportions  $p_i$  in the simplex, is given by

$$E[H] = \frac{1}{\ln(n+1)} \left[ \Psi(n+1) + \gamma - 1 + \frac{1}{n+1} \right].$$

This expectation tends to 1 as  $n \rightarrow \infty$ , as it is easily seen from the inequalities  $\ln n \leq \Psi(n+1) \leq \ln(n+1)$ .

## 2.2 The distribution of $A$

For  $A$  it is possible, on the contrary, to deduce an analytical formula for its probability distribution, because it is a simple linear function of the proportions  $p$ .

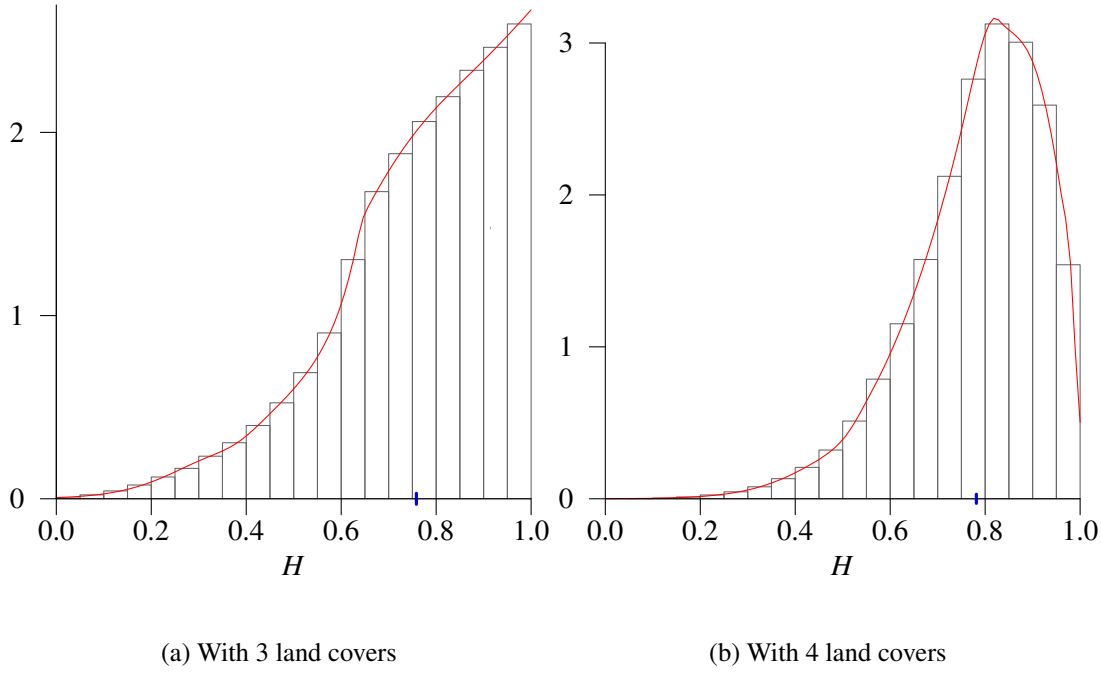


Figure 2: Histogram and density approximation for the random Shannon index  $H$  for 3 and 4 land covers. The blue line corresponds to the mean of the sample data.  $H$  has been calculated from a simulated uniform sample of  $p$  of size  $10^6$ .

Without loss of generality, we can assume that the weights  $w = (w_1, \dots, w_{n+1})$  are sorted and different:  $0 < w_1 < \dots < w_{n+1}$ . We can write

$$\begin{aligned} A &= \sum_{i=1}^n w_i p_i + w_{n+1} \left(1 - \sum_{i=1}^n p_i\right) \\ &= w_{n+1} - \sum_{i=1}^n s_i p_i, \end{aligned}$$

where  $s_i := w_{n+1} - w_i$ , and clearly  $0 < s_n < s_{n-1} < \dots < s_1 < w_{n+1}$ .

To obtain the distribution of  $A$  when  $p$  is uniform on  $\Delta'$ , let us compute first the probability density of  $\sum_{i=1}^n s_i p_i = w_{n+1} - A$ . We use a change of variable by means of the bijective linear transformation  $T: \Delta' \rightarrow B \subset \mathbb{R}^n$  given by

$$\begin{cases} v_1 = \sum_{i=1}^n s_i p_i \\ v_j = s_j p_j, \quad j = 2, \dots, n \end{cases}$$

where

$$B = \left\{ v \in \mathbb{R}^n : \sum_{i=1}^n \frac{v_i}{s_i} - \sum_{i=2}^n \frac{v_i}{s_1} \leq 1, \sum_{i=2}^n v_i \leq v_1, \text{ and } v_i \geq 0 \right\}.$$

The inverse mapping  $T^{-1}: B \rightarrow \Delta'$  is defined by

$$\begin{cases} p_1 = \frac{1}{s_1} (v_1 - \sum_{i=2}^n v_i) \\ p_j = s_j v_j, \quad j = 2, \dots, n \end{cases}$$

with Jacobian determinant equal to  $\prod_{i=1}^n \frac{1}{s_i}$ . Therefore, the density of the vector  $v = (v_1, \dots, v_n)$  is given by

$$f(v_1, \dots, v_n) = \begin{cases} n! \prod_{i=1}^n \frac{1}{s_i} & \text{if } v \in B \\ 0 & \text{otherwise} . \end{cases} \quad (7)$$

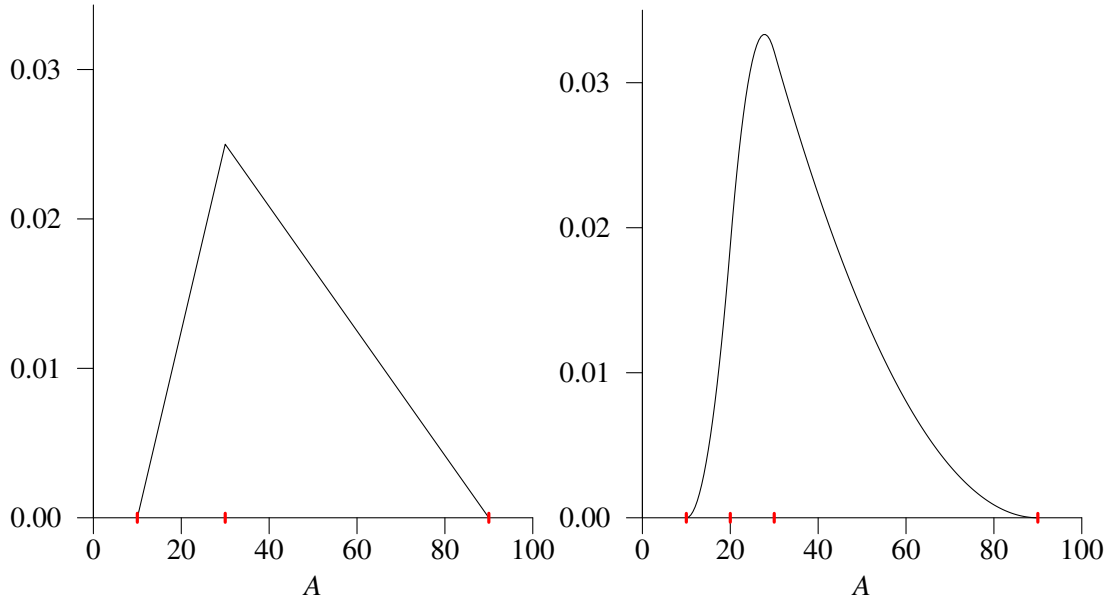
(a) With 3 land covers and  $w = (10, 30, 90)$ (b) With 4 land covers and  $w = (10, 20, 30, 90)$ 

Figure 3: Density of the appropriation  $A$  for 3 and 4 land covers, and for a particular vector of weights  $w$ , with values indicated by the red marks.

To obtain the density of  $v_1$ , we integrate (7) with respect to  $v_2, \dots, v_n$ . For fixed  $v_1, \dots, v_{k-1}$ , the variable  $v_k$  ranges from 0 to  $m_k$ , with

$$m_k = \min \left\{ v_1 - \sum_{i=2}^{k-1} v_i, \frac{s_1 s_k}{s_1 - s_k} \left( 1 - \frac{v_1}{s_1} - \sum_{i=2}^{k-1} v_i \frac{s_1 - s_i}{s_1 s_i} \right) \right\}.$$

Hence,

$$f(v_1) = \int_0^{m_2} \cdots \int_0^{m_n} n! \prod_{i=1}^n \frac{1}{s_i} dv_n \cdots dv_2,$$

which can be exactly computed for given values of  $s_1, \dots, s_n$ .

Finally, the density function of  $A$  is simply

$$f_A(a) = \begin{cases} f(w_{n+1} - a) & \text{if } a \in [w_1, w_{n+1}] \\ 0 & \text{otherwise.} \end{cases}$$

The graph of this function of  $a$  is depicted in Figure 3 for three and four land covers and some given values of  $w$ .

The expected value of  $A$  is easily computed using (5) directly, or reasoned by symmetry:

$$E[A] = \frac{1}{n+1} \sum_{i=1}^{n+1} w_i.$$

### 2.3 Expected value of $H$ for a given appropriation

We show in this subsection that a closed formula can be derived for the expected value of the Shannon index  $H$  conditioned to a given level of appropriation  $A$ . Specifically, we want to compute the function

$$a \mapsto E[H \mid A = a] \quad (8)$$

Since both  $H$  and  $A$  are functions of the vector of land covers  $p = (p_1, \dots, p_n) \in \Delta'$ , the conditional expectation can be computed by means of the conditional law of  $p$  given  $A(p) = a$ .



**Lemma.** Let  $X = (X_1, \dots, X_n)$  be a random vector following a continuous uniform distribution with support on a Borel set  $\Gamma \subset \mathbb{R}^n$  and let  $Y := \alpha_0 + \alpha_1 X_1 + \dots + \alpha_n X_n$ , for some constants  $\alpha_i \in \mathbb{R}$ .

Then, the conditional distribution of  $X$  given  $\{Y = a\}$  is uniform in  $\mathbb{R}^{n-1}$  with support on the intersection  $I_a := \Gamma \cap \{\alpha_0 + \alpha_1 x_1 + \dots + \alpha_n x_n = a\}$ , for almost all  $a$  with respect to the law of  $Y$ .

The fact stated in the lemma looks intuitive and it is indeed straightforward to prove. Notice, however, that the fact that  $\{\alpha_0 + \alpha_1 x_1 + \dots + \alpha_n x_n = a\}$  is a bundle of parallel lines is crucial, and that the result does not say anything about a particular value  $a$ , but should be understood with respect to the set of values  $a$  as a whole.

We apply the lemma to  $X = (p_1, \dots, p_n)$ ,  $\Gamma = \Delta'$ , and  $Y = A = w_{n+1} - \sum_{i=1}^n (w_{n+1} - w_i) p_i$ .

The intersection of the simplex  $\Delta'$  with the line  $\{A = a\}$  is given by

$$I_a = \{(p_1, \dots, p_{n-1}) \in \mathbb{R}^{n-1} : m_{k,a} \leq p_k \leq M_{k,a}, \forall k\},$$

where

$$m_{k,a} := \max \left\{ 0, \frac{w_{k+1} - a - \sum_{i=1}^{k-1} (w_{k+1} - w_i) p_i}{w_{k+1} - w_k} \right\},$$

$$M_{k,a} := \frac{w_{n+1} - a - \sum_{i=1}^{k-1} (w_{n+1} - w_i) p_i}{w_{n+1} - w_k}.$$

Taking into account that, on  $I_a$ , we can write  $p_n$  and  $p_{n-1}$  as a function of the other coordinates, namely,

$$p_n = \frac{w_{n+1} - a - \sum_{i=1}^{n-1} (w_{n+1} - w_i) p_i}{w_{n+1} - w_n}$$

and

$$p_{n+1} = 1 - \sum_{i=1}^n p_i = \frac{a - w_n + \sum_{i=1}^{n-1} p_i (w_n - w_i)}{w_{n+1} - w_n},$$

we have that the conditional expectation (8) is in fact a function of  $n - 1$  coordinates of  $p$ , and can be expressed as

$$\mathbb{E}[H \mid A = a] = \int_{I_a} C_a^{-1} \left[ -\sum_{i=1}^{n+1} p_i \log_{n+1} p_i \right] dp, \quad (9)$$

where

$$C_a := \int_{m_{1,a}}^{M_{1,a}} \dots \int_{m_{n-1,a}}^{M_{n-1,a}} dp_{n-1} \dots dp_1$$

is the volume of  $I_a$ .

The integral (9) can be computed exactly as a piecewise function that depends on the value of  $a$ . The result is given in Figures 4 i 5, for  $n + 1 = 3$  and  $n + 1 = 4$  and two sets of weights  $w$ . In all cases and dimensions the function (8) is continuous, piecewise concave, and non-smooth at the points  $w_i$ .

### 3 Shannon index and appropriation with real data

For the sake of simplicity, in this section we change  $n + 1$  to  $n$  and hereinafter the simplex will be

$$\Delta = \{(p_1, \dots, p_n) \mid p_i \geq 0, p_1 + \dots + p_n = 1\}.$$

Given a wide region, divided in small cells, the proportion of land covers in each cell will rarely be well represented by the uniform distribution. Not only some land covers can take more surface than others in the region, but also not all covers will be present in all cells.

To apply the method of the previous section with sample data, we need first to estimate from the data the probability distribution of land covers for the target region. Then, a large sample will be drawn from that distribution, and the conditional expectation  $\mathbb{E}[H \mid A = a]$  will be estimated from that sample.

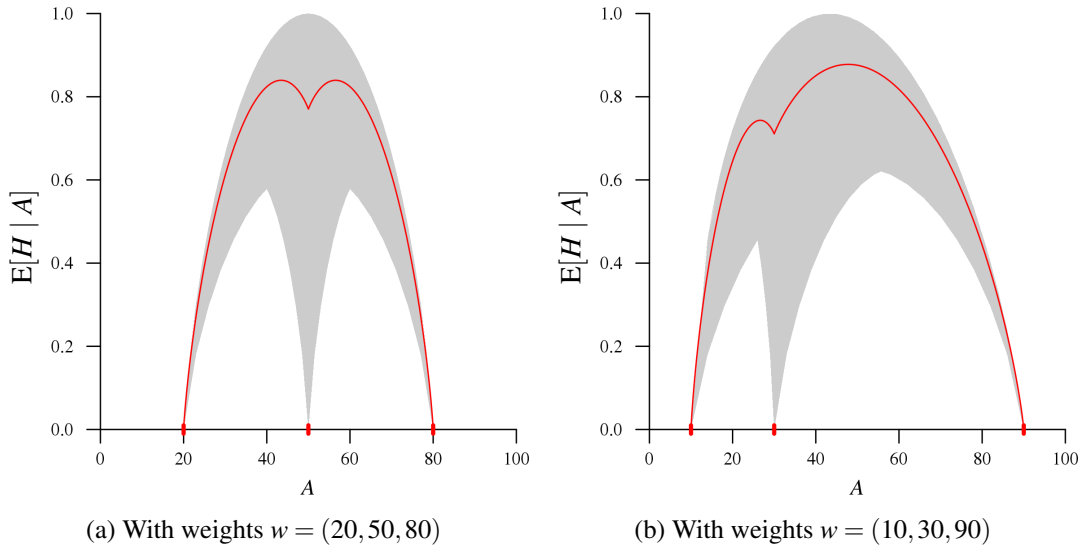


Figure 4: The red curve is the expected value of the Shannon index  $H$  as a function of the human appropriation  $A$ , for  $n + 1 = 3$  covers. The shaded area corresponds to the set of possible pairs of values  $(A, H)$ , and has been drawn by simulating one million points from its joint probability distribution.

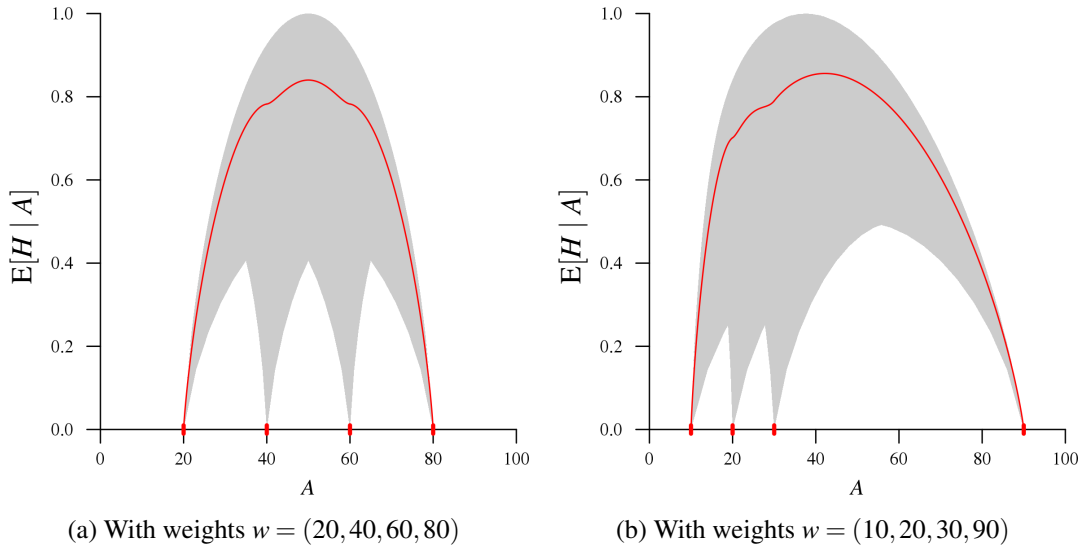


Figure 5: The analogues of Fig. 4 for  $n + 1 = 4$  covers, and the indicated weights. Note that the set of possible points and the conditional expectation curves are symmetrical if the weights are equidistant.

The analytical exact computation is of course no longer possible, since there is no a closed analytical expression for the distribution of  $A$ , unlike the uniform case. However, the estimated distribution of the proportions  $p$  allows to simulate as many values of  $H$  and  $A$  as desired, and these in turn allow to approximate  $E[H | A = a]$ . The quality of the result depends on the quality of the estimation of the distribution of  $p$  and on the number of values simulated.

This programme has some difficulties, that will be addressed in different subsections below. First, we develop the estimation of a density on a simplex by means of Dirichlet kernels. This estimation has numerical difficulties, that we solve in the second subsection. Next, we consider the global sampling strategy, taking into account the many points that lie in the facets of the simplex, which are themselves simplices of lower dimensions. Finally, we explain our procedure to choose the bandwidth parameter of the kernels, an important detail that will be postponed in the first subsection.

An option to avoid the difficulties with de Dirichlet kernels is to employ the log-ratios  $y_i = \log(p_i/p_{n+1})$ , see [1], or symmetric and isometric log-ratios, see [3], and then use kernels with unbounded domain, but these methods have serious drawbacks with samples whose points can very well be on the boundary of the simplex, as is in our case.

### 3.1 Kernel density estimation on the simplex

The estimation of probability distributions from data can be done in two ways: Either postulating a parametric family of distributions and estimating the parameters from the data, or by letting the data directly shape the distribution. In the second case, a probability *density function* is usually assumed to exist, and we speak of *non-parametric density estimation*.

We dismissed the first method due to the following reason: the only standard family of distributions with bounded support is the Dirichlet family, but we found that our data was far from being well represented by any of its members. Nevertheless we will use the Dirichlet family in a different way, as kernels to apply the *kernel density estimation* method. For the reader convenience, we recall here the definition of the Dirichlet family and the kernel method:

The density function of the Dirichlet distribution of dimension  $n > 1$  and positive parameters  $\alpha = (\alpha_1, \dots, \alpha_n)$  is

$$f(x_1, \dots, x_n) = \frac{1}{B(\alpha)} \prod_{j=1}^n x_j^{\alpha_j-1}, \quad (10)$$

supported by the simplex  $\Delta$ , where  $B$  is the multivariate Beta function:

$$B(\alpha) = \frac{\prod_{j=1}^n \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^n \alpha_j)}, \quad \text{and} \quad \Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx.$$

The kernel method, in general, consists of estimating the true density function  $f$  by

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^N K(x, z_i, \Lambda),$$

where  $K$  is the *kernel function*, which is a probability density function in  $x$  depending on the sample points  $z_i$ ,  $i = 1, \dots, N$ , and on an  $n \times n$  symmetric and positive-definite matrix  $\Lambda$ , called the *smoothing* or *bandwidth* matrix. As a function of  $x$ ,  $K$  attains its maximum at  $x = z_i$ . Parameters outside the diagonal in  $\Lambda$  define the degree of covariance between the kernel marginal laws, and the size of its eigenvalues are related to the kernel spread, that is, the greater the eigenvalues, the larger the spread in the corresponding eigenvector direction. In general, the kernel methods have good asymptotic properties.

In the absence of any relevant additional information, we will take  $\Lambda$  as a diagonal matrix with the same variance  $\lambda$  in all coordinate directions, and in consequence the kernel will be the Dirichlet density (10) with

$$\alpha_j = 1 + \frac{z_{ij}}{\lambda},$$

where  $z_{ij}$  is the  $j$ -th coordinate of  $z_i$ .

Using a kernel supported on the simplex  $\Delta$  ensures that the estimation is also supported on  $\Delta$ . The choice of the bandwidth parameter  $\lambda$  is crucial for an accurate estimation of the density. We have spent a considerable effort to get it right, and this is the contents of Subsection 3.4.

According with the assumptions above, our estimated density of the proportions  $p$  is given by

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^N \frac{\Gamma(n + \frac{1}{\lambda})}{\prod_{j=1}^n \Gamma(1 + \frac{z_{ij}}{\lambda})} \prod_{j=1}^n x_j^{z_{ij}/\lambda}. \quad (11)$$

As we will see, in the search of the optimal value of  $\lambda$ , we will need to evaluate (11) with  $\lambda$  in the order of  $10^{-3}$ . That means, the gamma functions in both numerator and denominator will have a very

large argument, with a subsequent loss of precision. For that reason, in Subsection 3.2 we look for an approximation of the gamma function to simplify the quotient before evaluating each part.

All of the above can be applied under the assumption that there exists a density on the simplex. In our case this is in fact not true, because there are data points in the lower dimensional facets of the simplex, corresponding to the cells on which not all land covers are present. We explain the solution in Subsection 3.3.

### 3.2 Numerical approximation of the estimated density

To get an appropriate numerical approximation of the quotient of gammas in (11), we use Weierstrass' formula

$$\Gamma(t+1) = e^{-\gamma} \prod_{k=1}^{\infty} (1+t/k)^{-1} e^{t/k},$$

where  $\gamma$  is the Euler-Mascheroni constant again. Denoting

$$C := \prod_{j=1}^{n-1} (n-j+\frac{1}{\lambda}),$$

the quotient in (11) can be written

$$\begin{aligned} \frac{\Gamma(n+\frac{1}{\lambda})}{\prod_{j=1}^n \Gamma(1+\frac{z_{ij}}{\lambda})} &= C \prod_{j=1}^n \frac{\Gamma(1+\frac{1}{\lambda})^{1/n}}{\Gamma(1+\frac{z_{ij}}{\lambda})} \\ &= C \prod_{j=1}^n \left[ e^{-\frac{\gamma}{\lambda}(\frac{1}{n}-z_{ij})} \prod_{k=1}^{\infty} \frac{1+\frac{z_{ij}}{k\lambda}}{(1+\frac{1}{k\lambda})^{\frac{1}{n}}} e^{\frac{1}{k\lambda}(\frac{1}{n}-z_{ij})} \right] \\ &= C \prod_{j=1}^n \left[ e^{-\frac{\gamma}{\lambda}(\frac{1}{n}-z_{ij})} \exp \left\{ \sum_{k=1}^{\infty} \left[ \frac{\frac{1}{n}-z_{ij}}{k\lambda} + \log \frac{1+\frac{z_{ij}}{k\lambda}}{(1+\frac{1}{k\lambda})^{\frac{1}{n}}} \right] \right\} \right]. \end{aligned}$$

Now we replace the series by a finite sum, with a controlled error, by means of the Euler–MacLaurin formula. Denoting by  $g(k)$  the expression in the internal square brackets (which depends also on  $z_{ij}$ ),

$$\sum_{k=m}^{\infty} g(k) = \int_m^{\infty} g(x)dx + \frac{1}{2}g(m) - \sum_{r=1}^s \frac{B_{2r}}{(2r)!} g^{(2r-1)}(m) + R_s,$$

with the remainder term satisfying

$$|R_s| \leq \frac{|B_{2s+2}|}{(2s+2)!} |g^{(2s+1)}(m)|,$$

and where  $B_r$  are the Bernoulli numbers, that can be defined recursively as

$$B_r = - \sum_{k=0}^{r-1} \frac{n!B_k}{k!(r+1-k)!}, \quad B_0 = 1.$$

The formula is true under the conditions

- (i)  $g^{(2s+2)}(x)g^{(2s+4)}(x) > 0$ , for  $x \in [m, \infty]$ ,
- (ii)  $\lim_{x \rightarrow \infty} g^{(2s+1)}(x) = 0$ .

If we call  $\tilde{f}$  the approximation of the estimated density  $\hat{f}$  when disregarding the remainder  $R_s$ , and  $M := \exp\{\max_{i,j} |R_s|\}$ , then

$$\tilde{f}M^{-n} \leq \hat{f} \leq \tilde{f}M^n.$$

Thus to obtain a final relative error  $\eta$ , we have to find an  $\varepsilon$  such that  $\varepsilon \geq \max_{i,j} |R_s|$  and  $\exp\{n\varepsilon\} \leq (1 + \eta)$ . This amounts to take

$$\varepsilon = \frac{1}{n} \log(1 + \eta) ,$$

and to find natural numbers  $s$  and  $m$  such that  $\max_{i,j} |R_s| < \varepsilon$ , and satisfying the conditions of the Euler-MacLaurin formula. In this way, we will finally get the approximation

$$\begin{aligned} \bar{f}(x) = \frac{C}{N} \sum_{i=1}^N \exp \left\{ \sum_{j=1}^n \left[ \frac{1}{\lambda} \left( -\gamma \left( \frac{1}{n} - z_{ij} \right) + z_{ij} \log(x_j) \right) \right. \right. \\ \left. \left. + \sum_{k=1}^{m-1} g(k) + \int_m^\infty g(x) dx + \frac{1}{2} g(m) \right. \right. \\ \left. \left. - \sum_{r=1}^s \frac{B_{2r}}{(2r)!} g^{(2r-1)}(m) \right] \right\} , \end{aligned}$$

with

$$(1 + \eta)^{-1} \leq \hat{f}/\bar{f} \leq (1 + \eta) .$$

The conditions to apply the Euler-MacLaurin formula are in our case always fulfilled for very small integers  $m$  and  $s$ , when taking  $\eta = 10^{-4}$ . The minimal ones are readily found by simple search.

### 3.3 Sampling strategy

Our real dataset contains many cells in which one or more land covers are not present. Hence, the theoretical distribution from which they are taken does not actually possess a density on the simplex  $\Delta$ . However, we can assume the existence of a density on the subsimplices obtained by restricting some of the coordinates to be zero. Indeed, the resolution of our data is sufficient to estimate the density on each subsimplex, using the points that lie on it, except in a few cases.

If  $f_\delta$  is the theoretical density on the subsimplex  $\delta$ , and  $q_\delta := P\{p \in \delta\}$  is the theoretical probability that one random point of  $\Delta$  lie on the subsimplex  $\delta$ , the overall probability distribution can be described as

$$\begin{aligned} P\{p \in A\} &= \sum_{\delta} q_\delta \cdot P\{p \in A \cap \delta \mid p \in \delta\} \\ &= \sum_{\delta} q_\delta \cdot \int_{A \cap \delta} f_\delta(x) dx , \end{aligned}$$

for any Borel set  $A \subset \Delta$ , and where the sum runs over all subsimplices.

To estimate the distribution of the whole dataset we can therefore proceed in the following way: The probabilities  $q_\delta$  can be estimated by the sample proportion  $\hat{q}_\delta$  of points lying in  $\delta$ ; the densities on each subsimplex  $\delta$  can be estimated and approximated as  $\bar{f}_\delta$  by the method just described on Subsection 3.2. One obtains the estimate

$$P\{p \in A\} \approx \sum_{\delta} \hat{q}_\delta \cdot \int_{A \cap \delta} \bar{f}_\delta(x) dx .$$

Although there is no explicit form for the densities  $\bar{f}_\delta$ , we can evaluate them at arbitrary points  $x$  and apply the acceptance/rejection method to simulate a large sample following this distribution. Specifically:

1. Choose randomly a subsimplex  $\delta$  with probability  $\hat{q}_\delta$ .
2. Generate a random vector  $x$  with uniform distribution on  $\delta$ , with the method of Section 2.

3. Generate a random number  $u$  with uniform distribution on  $[0, 1]$  and evaluate

$$uC_\delta \leq \tilde{f}_\delta(x) .$$

If the inequality holds true, accept  $x$  as a new point of the sample; otherwise, reject it and go back to step 2.

4. Go back to step 1 until the desired sample size is reached.

In step 3,  $C_\delta$  is any constant satisfying  $C_\delta \geq \max\{\tilde{f}_\delta(x)\}$ . Ideally, this constant must be an upper bound as tight as possible of the density function  $\tilde{f}_\delta$ , in order not to reject too many generated points. However, we only know this density in a big, but finite, number of points. If, during the run of the acceptance/rejection method, a value of  $\tilde{f}_\delta$  greater than the chosen  $C_\delta$  is found, then some of the already accepted points must have been actually rejected. From the practical point of view, we have preferred in our case study to take a safe upper bound, so that none of the accepted points have to be discarded later, despite the larger running times incurred.

The absolute error in the probability of accepting a point  $x$  based in the approximate density  $\tilde{f}$  in step 2 above, when it would have been rejected if  $\hat{f}$  could be used, it is bounded by the constant  $\eta$ . Indeed, the difference in the probabilities to accept the point in the two cases is

$$\begin{aligned} 0 &\leq \frac{1}{C_\delta} (\tilde{f}(x) - \hat{f}(x)) \leq \frac{1}{C_\delta} (\tilde{f}(x) - \tilde{f}(x)(1 + \eta)^{-1}) \\ &= \frac{\tilde{f}(x)}{C_\delta} (1 - (1 + \eta)^{-1}) \leq 1 - \frac{1}{1 + \eta} \leq \eta . \end{aligned}$$

Analogously, one can show that the difference in the probability of rejecting a point is less than the same constant  $\eta$ .

### 3.4 Choosing the bandwidth parameter

As mentioned before (see Subsection 3.1) the goodness of the estimation of a density by a kernel method depends heavily on the choice of the bandwidth (or smoothing) parameter  $\lambda$ . In general, the larger the sample size, the smaller the bandwidth should be, or, in other words, the less influence each sample point must have on the final estimation.

In our case, the initial sample size is  $N = 3360$ . Although we have to work independently on each subsimplex, the bandwidths will tend to be small anyway, as this is what creates the numerical problem that we have addressed in Section 3.2.

In the frequently cited paper by [9], and in [1], the authors propose to choose the smoothing parameter  $\lambda$  that maximises the pseudo-likelihood

$$\prod_{i=1}^N \frac{1}{N-1} \sum_{j \neq i} K(x_i, x_j, \lambda I) ,$$

where  $x$  are the sample points,  $N$  is the sample size and  $I$  is the identity matrix.

Instead, we will adjust  $\lambda$  according to the use that we will make of the estimated density. Namely, we want to approximate the function that maps appropriation levels to the conditional expectation of the Shannon index given that level:

$$a \mapsto E[H \mid A = a] . \quad (12)$$

To this end, we proceed with the following steps, on each subsimplex:

1. Assume the points in the subsimplex follow a Dirichlet distribution.
2. Estimate the parameters  $\alpha$  of the distribution (10). We have used the maximum likelihood method implemented in the function `dirichlet.mle` of the R package `sirt`.

3. Generate a large number of points (e.g.  $10^6$ )  $Y$  with the estimated distribution. These data plays the role of 'synthetic population' in this process.
4. Sample a subset  $Z$  of  $Y$  of the same size as the part of the real sample that lies on the subsimplex. These data  $Z$  is used as the 'synthetic sample' for the next steps.
5. For a given value of  $\lambda$ , apply the procedure explained in 3.3 to simulate a sample  $X_\lambda$  of the estimated density (say, of size  $10^4$ ).
6. Measure the fit of the simulated data with the 'synthetic population'  $Y$  using the integrated square error

$$\int_{w_1}^{w_n} (\phi_Y(a) - \phi_{X_\lambda}(a))^2 da, \quad (13)$$

where  $\phi_Y$  and  $\phi_{X_\lambda}$  are the functions (12) for  $\phi$  corresponding respectively to the population  $Y$ , and to the sample  $X_\lambda$ .

7. Repeat steps 5–6 to choose  $\lambda$  that minimises (13).

Some remarks are in order about the scheme above:

- a) In our case study, it is possibly not true that the data can be well represented by a Dirichlet distribution; if we knew it were, then we would be better off adopting directly the density that results from the maximum likelihood estimate. However, we use it at this point as a proxy because of its support on the simplex, and only to obtain a plausible bandwidth; using the uniform distribution on the subsimplices for the same purpose will be even more inadequate.
- b) The sample sizes of  $Y$  and  $X_\lambda$  are arbitrary. They should simply look like a (big) population and an (also big) sample from it. On the contrary, we think that it is realistic to make the size of  $Z$  equal to the size of the real data at hand. The integral in (13) cannot be computed exactly, because the function (12) cannot be either. We discretise the values of  $A$  to obtain a stepwise approximation of  $\phi$ , so that the integral is in fact approximated by a finite sum. But this is fine, since the final result will necessarily be given as a discretised function.
- c) Finally, the integrated square error is not the only possible criterion for the choice of  $\lambda$ ; others can be used, depending on the application sought.

## 4 Results

In this section we present the results of the procedures proposed in Sections 2 and 3 when applied to the data of the case study described in the introduction. All figures referenced have been grouped together at the end of the paper, for easy comparison.

The four types of land covers are: the semi-natural land covers, with lowest human intervention (forest, scrubland, prairie and bedrock, and wetland),  $p_1$ ; the cropland, both irrigated and dry crops,  $p_2$ ; the land covers with groves,  $p_3$ ; and the urban and industrial surfaces,  $p_4$ .

This grouping has been established according to the similarity in the weights of the original ten land covers, the latter taken from [15], and each type is assigned the mean of the original weights (see Table 1). There are different values for each year, due to the changes in the exploitation of land covers over time. From 1956 to 2000 there is a general reduction in the values of  $w$ . It is known that in the last decades of the twentieth century there has been in Mallorca a progressive abandonment of the arable land, inducing an expansion of forests, from which humans extract little profit [15].

The real data is distributed in subsimplices as described in Table 2. As we can see there, the dominant subsimplex in 1956 and 1973 is the one comprising 'semi-natural', 'cropland' and 'groves' covers. Such combinations are usually referred as *mosaic landscapes*. Their frequency clearly declines in 2000, where the combination of 'semi-natural' and 'cropland' prevails.

year	$w_1$	$w_2$	$w_3$	$w_4$
1956	51.042	78.880	89.993	95.730
1973	43.958	76.200	85.322	94.792
2000	48.542	74.978	81.837	93.958

Table 1:  $w$  values for each year.

In Figure 6, a scatter plot of the joint values of  $H$  and  $A$  is shown, for each of the three times periods (1956, 1973 and 2000). Recall that the support of the feasible pairs has the irregular greyed shape that we saw in Figure 5, with the ‘legs’ of the region resting over the weight values in the horizontal axis; hence the white empty zones in the scatter plot. Dots are plotted with some degree of transparency; the apparently solid lines describing arcs between the legs are points whose corresponding proportions  $p$  lie in the edge joining two vertices of the simplex. Some of these edges are more populated than others, or more evenly distributed, and those arcs are therefore more noticeable in the figure.

In Figure 7, the same scatter plot of the pairs  $(A, H)$  is depicted, for the enlarged dataset obtained by the sampling method of Subsection 3.3, and the three corresponding time periods. Table 2 shows the  $\lambda$  values on each subsimplex obtained following the optimisation procedure of Subsection 3.4. Of course, vertices of the simplex does not have a density. Also, we have not estimated a density for subsimplices with less than 30 data points; instead, we have sampled them as a discrete equally probable population. The threshold of 30 is arbitrary.

Except for the number of points, Figures 6 and 7 look indeed quite similar, which speaks in favour of our method of estimation of the probability distribution of the proportions in the simplex. To reinforce this impression, in Figures 8 and 9 we compare estimations of the join density of  $A$  and  $H$  both from the initial data and for the enlarged sample. In these figures we have used a simple Gaussian kernel density estimation in the plane, just to have a visual quick idea of the similarities between the large synthetic sample and the original one, in order to validate the whole computation of the conditional expectations in Section 3.

Subsimplices $\delta$	1956		1973		2000	
	$N_\delta$	$\lambda$	$N_\delta$	$\lambda$	$N_\delta$	$\lambda$
1 0 0 0	228	-	224	-	226	-
0 1 0 0	30	-	27	-	240	-
1 1 0 0	109	0.007	98	0.029	1094	0.001
0 0 1 0	84	-	78	-	24	-
1 0 1 0	787	0.003	766	0.002	199	0.039
0 1 1 0	489	0.013	454	0.026	212	0.009
1 1 1 0	1311	0.006	1208	0.006	532	0.004
0 0 0 1	1	-	3	-	7	-
1 0 0 1	3	-	12	-	28	-
0 1 0 1	8	-	14	-	144	0.008
1 1 0 1	8	-	13	-	298	0.007
0 0 1 1	39	0.027	51	0.05	24	-
1 0 1 1	59	0.032	111	0.015	29	-
0 1 1 1	105	0.014	141	0.011	136	0.015
1 1 1 1	99	0.035	160	0.03	167	0.031

Table 2: Subsimplex typologies  $\delta$ , corresponding to different combinations of land covers (1 indicates presence, 0 absence); size  $N_\delta$  of each subsimplex, and chosen values of  $\lambda$ .

#### 4.1 Shannon index conditioned to the appropriation

In Figure 10 we can see superimposed the plots of  $a \mapsto E[H \mid A = a]$ , with the assumptions of both Sections 2 and 3. The red curve is the analytic result obtained assuming a uniform distribution of covers,



whereas the blue points are the ones we have obtained with the real data of our case study and the procedure of Section 3. The vertical grey lines indicate the values  $w$ .

Real data produce, for all time periods and for practically all values of appropriation, an expected value of the Shannon index  $H$  lower than with the uniform distribution. This was absolutely expected, because in the real dataset rarely all types of cover appear in a single cell (only 99 over 3360 cases, see Table 2), nor the appearing ones look like evenly distributed. Recall that the Shannon index is maximal when all proportions coincide.

Concerning the annual evolution, figures show a strong similarity in the expected  $H$  for 1956 and 1973, whereas there are noticeable differences in 2000. First, there is a high decrease around  $a = w_2$ , motivated by the intensification of  $p_2$ . Secondly, the expectation after  $w_2$  increases due to the growth of urban areas combined with other land covers. The maximum of the expectation, in fact, jumps to the interval  $[w_3, w_4]$ .

## 4.2 Diversity and urban land cover

At the scale we are working in, one may consider that the urban cover is not a real habitat for living species (except humans). It has been proposed in [14] to use a variation of the Shannon index that penalises the presence of urban areas, as indicator of habitat diversity:

$$L := (1 - p_u) \left( - \sum_{i=1}^n p_i \log_n p_i \right),$$

where  $p_i$  are the proportions of non-urban covers, inside the total of non-urban surface, and  $p_u$  is the proportion of urban surface in the cell. With the obvious notation,

$$p_i = \frac{S_i}{S_{\text{Cell}} - S_u}, \quad p_u = \frac{S_u}{S_{\text{Cell}}}.$$

The maximum of  $L$  is 1 and it corresponds to  $p_i = \frac{1}{n}$ ,  $p_u = 0$ , with appropriation  $A = \frac{1}{n} \sum_{i=1}^n w_i$ .

As we pointed out before, there is no problem in applying the same methodology of Section 3 to  $L$  or other indices depending only on  $p$ . In Figure 11 one can see the relation we have obtained between the appropriation  $A$  and the conditional expectation  $E[L | A]$ . Again, the red curve corresponds to the expectation of the index  $L$  for each value of  $A$  when the land covers, including  $p_u$ , are uniformly distributed. In contrast with the case of  $H$ , this conditional expectation is in some intervals smaller than the values derived from the real data, represented by the blue dots.

The temporal evolution of  $E[L | A]$  in this figure reveals an evident change in the landscape structure from 1973 to 2000, essentially due to the urban growth and the decline of mosaic prevalence in that time interval. The changes can be partially explained with the help of Table 2.

First, we observe the much lower values around  $w_2$ : The number of cells where only the second type of cover is present (0100 in the table), or with the second and the urban cover (0101), have increased notably, and all of them produce a value  $L = 0$ . Therefore, the mean of the  $L$  index for values of appropriation in the interval  $[w_2, w_4]$  must be lower in 2000 than in 1973. This can be expected also by comparing the density of points in graphs (b) and (c) of Figure 6 or Figure 7, around  $w_2$ . However, on  $[w_3, w_4]$  this effect is more than compensated by the fact that covers of type 0011 have decreased; without the factor  $(1 - p_u)$ , which is small in this region, the graph will in fact get even higher, as in the case of  $E[H | A]$  in Figure 10. The effect of  $(1 - p_u)$  is almost negligible on  $[w_1, w_3]$ .

Concerning the interval  $[w_1, w_2]$ , in which the blue curve tends also to be lower in 2000 than in 1973 (most notably in the right half of the interval), a possible explanation is the smaller number of cells with the non-urban mosaic (1110), together with the increase of the cells of type (1100), two facts that are of course related. Indeed, this produces a higher proportion of values of  $L$  at the minimum possible value, on the arc joining  $w_1$  and  $w_2$  (the evolution is again apparent on Figures 6 and 7), and therefore a decrease of the expected value of  $L$  given  $A \in [w_1, w_2]$ .

While the temporal evolution of the dotted blue curve  $E[H | A]$  shows a land cover diversity loss for low values of the appropriation  $A$ , and a gain for high values of  $A$ , the curve based on index  $L$  helps to

analyse better, in our opinion, the effect of Mallorca urban expansion on habitat diversity. Namely, this effect is reflected in the general decrease of the values of the conditional expectation  $E[L | A]$  over the whole range of appropriation levels.

## 5 Computational notes

The computations have been done using R with the following setup:

- R version 3.3.1 (2016-06-21), x86\_64-pc-linux-gnu
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: CDM 5.0-0, knitr 1.13, logspline 2.1.9, mvtnorm 1.0-5, sirt 1.12-2, TAM 1.995-0

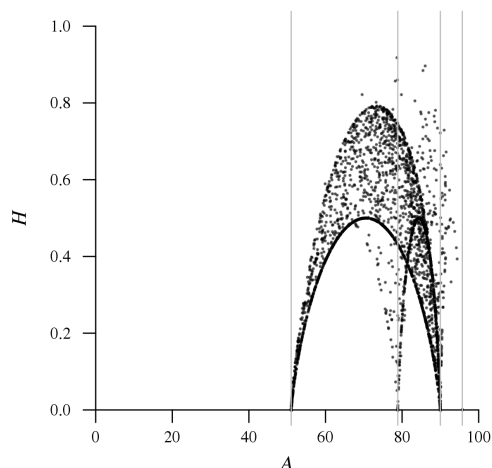
The C language has also been used in the most time-consuming routines.

The whole procedure of Section 3 is computationally intensive, due to the optimisation step to choose the right value of the parameter  $\lambda$  for each subsimplex, including an acceptance/rejection simulation for each tentative value. It is not a prohibitive load, though. The computational complexity is of course exponential as a function of the number of different covers, since there are  $2^{n+1} - 1$  subsimplices in the  $n$ -dimensional standard simplex  $\Delta$  in  $\mathbb{R}^{n+1}$ . For this reason, and to have enough sample size in most of the subsimplices, we grouped together the ten different covers of the original data into four classes of similar weights.

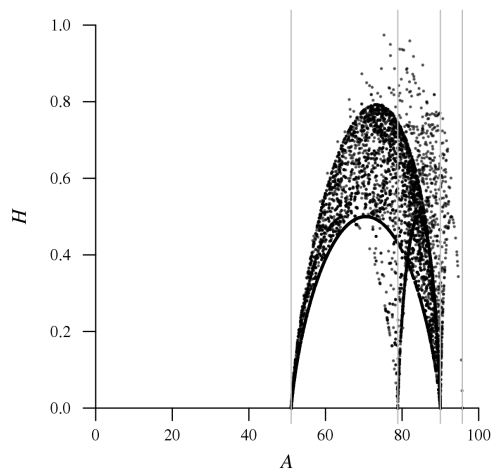
## References

- [1] J. Aitchison and I. J. Lauder. Kernel density estimation for compositional data. *Applied statistics*, pages 129–137, 1985.
- [2] A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, editors. *Compositional Data Analysis in the Geosciences*. Geological Society of London, 2006.
- [3] J. E. Chacón, G. Mateu-Figueras, and J. A. Martín-Fernández. Gaussian kernels for density estimation with compositional data. *Computers & Geosciences*, 37(5):702 – 711, 2011.
- [4] Robert K. Colwell. Biodiversity: Concepts, patterns, and measurement. In Simon A. Levin, Stephen R. Carpenter, H. Charles J. Godfray, Ann P. Kinzig, Michel Loreau, Jonathan B. Losos, Brian Walker, David S. Wilcove, and Christopher G. Morris, editors, *The Princeton Guide to Ecology*, chapter III.1, pages 257–263. Princeton University Press, 2009.
- [5] David J. Currie, Gary G. Mittelbach, Howard V. Cornell, Richard Field, Jean-Francois Guégan, Bradford A. Hawkins, Dawn M. Kaufman, Jeremy T. Kerr, Thierry Oberdorff, Eileen O’Brien, and J. R. G. Turner. Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecology Letters*, 7(12):1121–1134, 2004.
- [6] Jeremy W. Fox. The intermediate disturbance hypothesis should be abandoned. *Trends in Ecology & Evolution*, 28(2):86–92, 2013.
- [7] Kevin J. Gaston. Global patterns in biodiversity. *Nature*, 405(6783):220–227, May 2000.
- [8] GIST. Mapes de cobertes del sòl de les illes balears (1:25.000): 1956(1973), 1995, 2000. *Universitat de les Illes Balears, Departament de Ciències de la Terra, Grup d’Investigació de Sostenibilitat i Territori, Palma de Mallorca*, 2009.
- [9] J. D. E. Habbema, J. Hermans, and K. van den Broek. A stepwise discriminant analysis program using density estimation. In *COMPSTAT 1974: Proceedings in Computational Statistics*, pages 101–110. Physica-Verlag, 1974.
- [10] Helmut Haberl, K. Heinz Erb, Fridolin Krausmann, Veronika Gaube, Alberte Bondeau, Christoph Plutzer, Simone Gingrich, Wolfgang Lucht, and Marina Fischer-Kowalski. Quantifying and mapping the human appropriation of net primary production in earth’s terrestrial ecosystems. *Proceedings of the National Academy of Sciences*, 104(31):12942–12947, 2007.

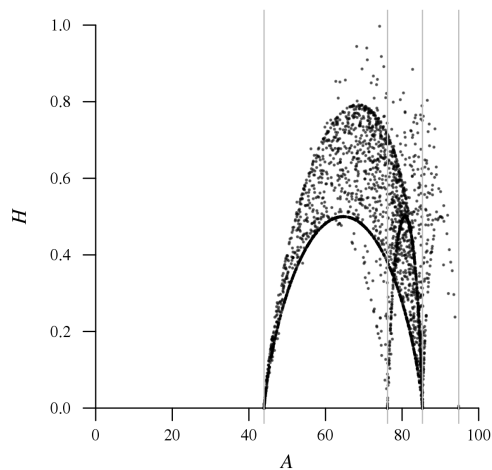
- [11] Helmut Haberl, Niels B. Schulz, Christoph Plutzer, Karl Heinz Erb, Fridolin Krausmann, Wolfgang Loibl, Dietmar Moser, Norbert Sauberer, Helga Weisz, Harald G. Zechmeister, and Peter Zulka. Human appropriation of net primary production and species diversity in agricultural landscapes. *Agriculture, Ecosystems & Environment*, 102(2):213 – 218, 2004.
- [12] Charles Kooperberg and Charles J Stone. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4):301–328, 1992.
- [13] Tom Leinster and Christina A. Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2006.
- [14] Joan Marull, Carme Font, Roc Padró, Enric Tello, and Andrea Panazzolo. Energy–landscape integrated analysis: A proposal for measuring complexity in internal agroecosystem processes (Barcelona Metropolitan Region, 1860–2000). *Ecological Indicators*, 66:30–46, 2016.
- [15] Joan Marull, Carme Font, Enric Tello, Nofre Fullana, Elena Domene, Manel Pons, and Elena Galán. Towards an energy–landscape integrated analysis? Exploring the links between socio-metabolic disturbance and landscape ecology performance (Mallorca, Spain, 1956–2011). *Landscape Ecology*, 31(2):317–336, 2015.
- [16] Joan Marull, Enric Tello, Nofre Fullana, Ivan Murray, Gabriel Jover, Carme Font, Francesc Coll, Elena Domene, Veronica Leoni, and Trejsi Decolli. Long-term bio-cultural heritage: exploring the intermediate disturbance hypothesis in agro-ecological landscapes (Mallorca, c. 1850–2012). *Biodiversity and Conservation*, 24(13):3217–3251, 2015.
- [17] Reuven Y Rubinstein and Benjamin Melamed. *Modern simulation and modeling*. Wiley, 1998.
- [18] J Tews, U Brose, V Grimm, K Tielbörger, MC Wichmann, M Schwager, and F Jeltsch. Animal species diversity driven by habitat heterogeneity/diversity: the importance of keystone structures. *Journal of Biogeography*, 31(1):79–92, 2004.
- [19] Peter M. Vitousek, Paul R. Ehrlich, Anne H. Ehrlich, and Pamela A. Matson. Human appropriation of the products of photosynthesis. *BioScience*, 36(6):368–373, 1986.



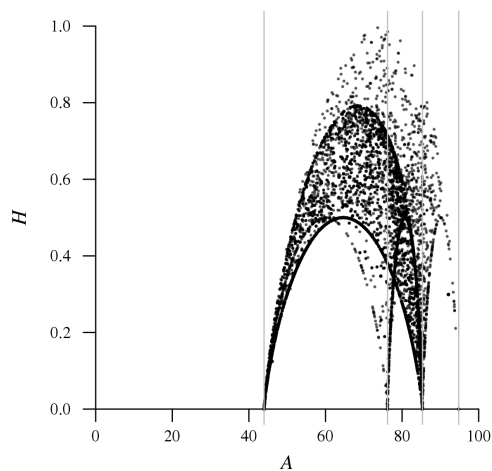
(a) 1956



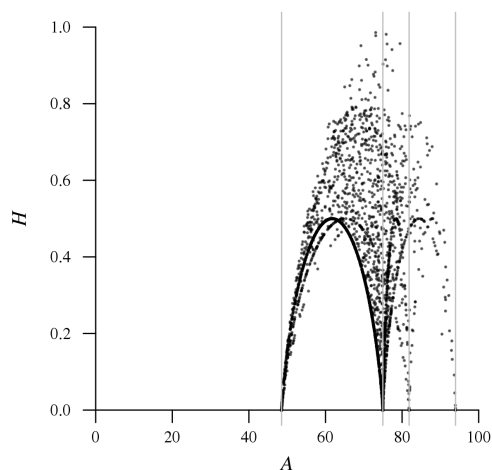
(a) 1956



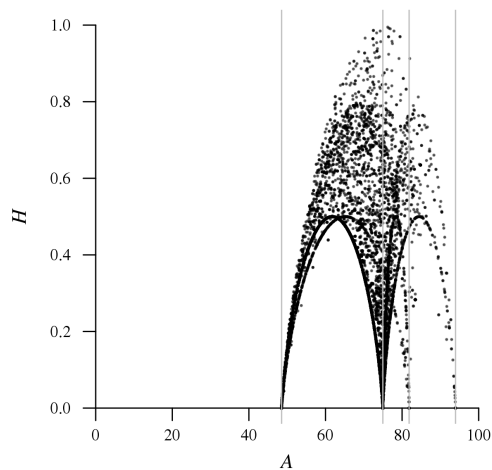
(b) 1973



(b) 1973



(c) 2000



(c) 2000

Figure 6: Values  $(A, H)$  of the real data from Mallorca Island (four land covers, 3360 points).

Figure 7: Simulated values  $(A, H)$  generated from the estimated distribution, and a sample size of  $10^4$  points.

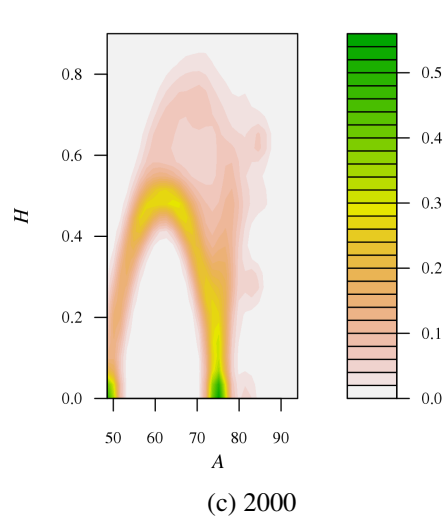
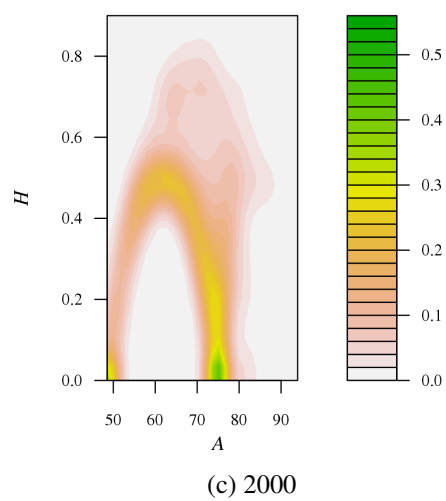
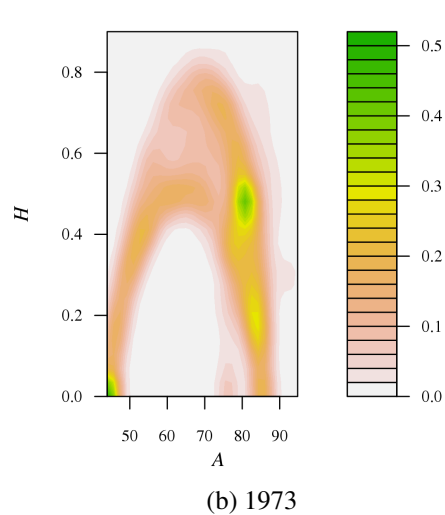
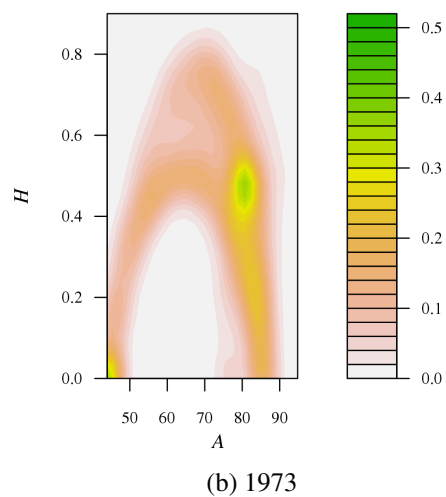
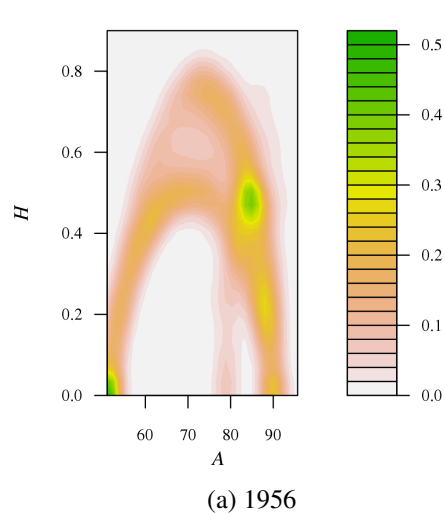
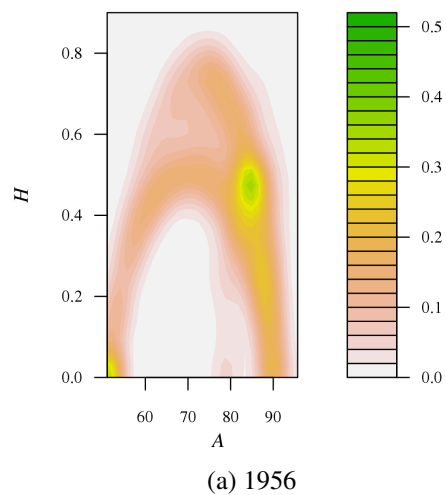
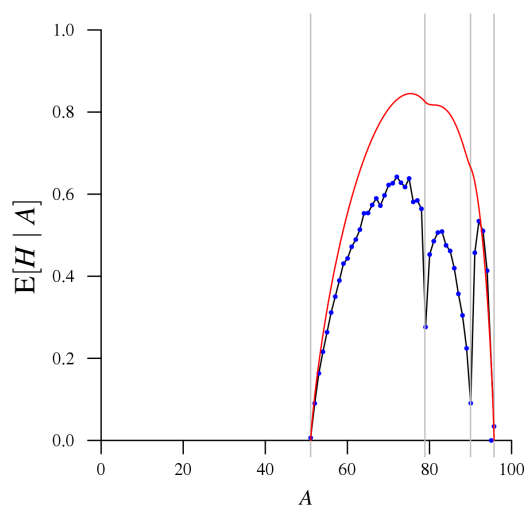
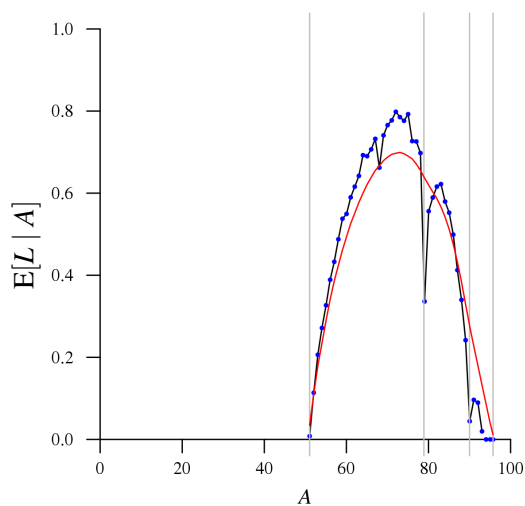


Figure 8: ‘Filled-contour plot’ of the two-dimensional of  $(A, H)$ , estimated from real data.

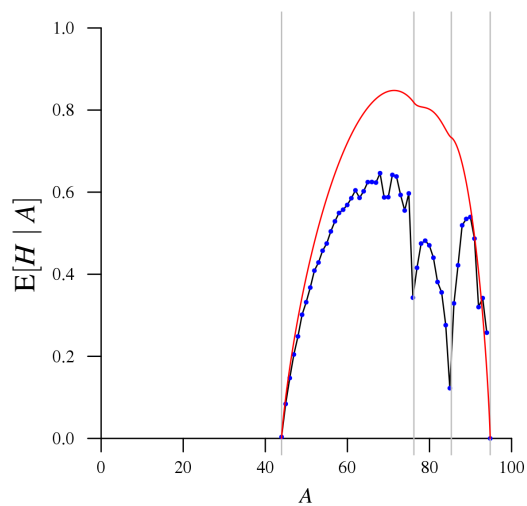
Figure 9: Analogue of Figure 8 for the simulated data.



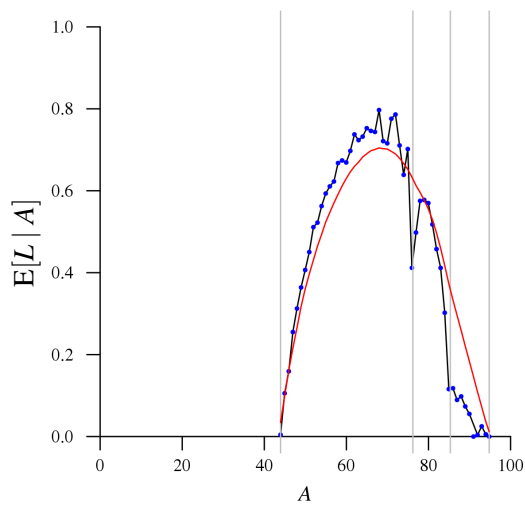
(a) 1956



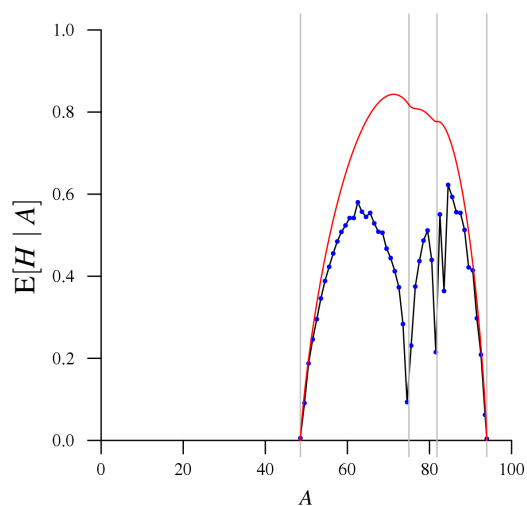
(a) 1956



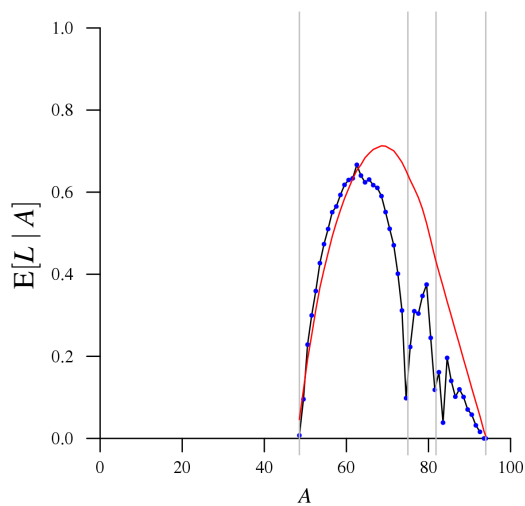
(b) 1973



(b) 1973



(c) 2000



(c) 2000

Figure 10: Conditional expectation  $E[H | A = a]$  with the uniform distribution of covers (red curve) and starting from the data of the case study (blue dots).

Figure 11: Analogue of Figure 10 with the indicator  $L$ .